Project Title	**Vaccination Data Analysis and Visualization**
Skills take away From This Project	Python script,data cleaning, EDA, SQL, Power BI
Domain	Public Health and Epidemiology

Problem Statement:

Analyze global vaccination data to understand trends in vaccination coverage, disease incidence, and effectiveness. Data will be cleaned, and stored in a SQL database. Power BI will be used to connect to the SQL database and create interactive dashboards that provide insights on vaccination strategies and their impact on disease control.

Business Use Cases:

- **Public Health Strategy:**
  - Assess the effectiveness of vaccination programs in different regions and populations.
  - Prioritize areas with low vaccination coverage for targeted interventions.
- **Disease Prevention:**
  - Identify diseases with high incidence rates despite vaccination efforts, suggesting vaccine inefficacies or areas for improvement.
  - Support policies on booster vaccines or new vaccine introductions.
- **Resource Allocation:**
  - Determine regions with low vaccination coverage and plan targeted resource distribution to improve vaccination rates.
  - Forecast vaccine demand based on current trends for better supply chain management.
- **Global Health Policy:**
  - Provide data-driven recommendations for vaccination policy formulation.
  - Support governments and health organizations with evidence on vaccine effectiveness.

Approach:

Data Cleaning.

- **Handle Missing Data**: Impute missing values or remove incomplete records.
- **Normalize Units:** Ensure consistency in units across datasets (e.g., percentage of coverage, number of reported cases).
- **Date Consistency:** Format date fields uniformly across tables for easier analysis.

SQL Database Setup

- **Create Tables:** Store the extracted and cleaned data into relational SQL tables (e.g., vaccination data, disease incidence data, antigen data).
- **Normalize Data:** Structure the data into separate tables (e.g., vaccines, diseases, countries, years) to avoid redundancy and improve querying performance.
- **Data Integrity:** Implement primary and foreign keys to ensure referential integrity.

Power BI Integration.

- **Connect Power BI to SQL Database:**
    - Use Power BI's SQL connector to link to the SQL database and pull in the relevant tables for analysis.
    - Set up scheduled refreshes for updated data.

Data Visualization in Power BI

- **Create Interactive Dashboards:**
    - Use Power BI to create dynamic and visually engaging reports with filters and slicers for users to explore the vaccination data.
    - Visualize vaccination rates, disease incidence, and antigen coverage over time and across regions.
- **Key Visualizations:**
    - **Geographical Heatmaps:** Display vaccination coverage and disease incidence by region.
    - **Trend Lines/Bar Charts:** Show trends in vaccination coverage, disease rates, and effectiveness over multiple years.
    - **Scatter Plots:** Correlate vaccination coverage and disease incidence across different countries or regions.
    - **KPI Indicators:** Track progress toward vaccination goals and health targets.

Exploratory Data Analysis (EDA):

- Analyze vaccination coverage, disease incidence trends, and regional disparities using statistical summaries and correlation analysis. Visualize insights with bar charts, heatmaps, and line graphs to identify patterns, highlight low-coverage areas, and assess the impact of vaccination on disease reduction.

Questions to be answered:

Answer the below questions through your analysis and visualizations:

Easy level:

1. How do vaccination rates correlate with a decrease in disease incidence?
2. What is the drop-off rate between 1st dose and subsequent doses?.
3. Are vaccination rates different between genders?
4. How does education level impact vaccination rates?
5. What is the urban vs. rural vaccination rate difference?
6. Has the rate of booster dose uptake increased over time?
7. Is there a seasonal pattern in vaccination uptake?
8. How does population density relate to vaccination coverage?
9. How do vaccination rates correlate with a decrease in disease incidence?
10. Which regions have high disease incidence despite high vaccination rates?

Medium level (combination of different tables):

1. Is there a correlation between vaccine introduction and a decrease in disease cases?
2. What is the trend in disease cases before and after vaccination campaigns?
3. Which diseases have shown the most significant reduction in cases due to vaccination?
4. What percentage of the target population has been covered by each vaccine?
5. How does the vaccination schedule (e.g., booster doses) impact target population coverage?
6. Are there significant disparities in vaccine introduction timelines across WHO regions?
7. How does vaccine coverage correlate with disease reduction for specific antigens?
8. Are there specific regions or countries with low coverage despite high availability of vaccines?
9. What are the gaps in coverage for vaccines targeting high-priority diseases (e.g., TB, Hepatitis B)?
10. Are certain diseases more prevalent in specific geographic areas?

Scenario based:

1. A government health agency wants to identify regions with low vaccination coverage to allocate resources effectively.
2. A public health organization wants to evaluate the effectiveness of a measles vaccination campaign launched five years ago.
3. A vaccine manufacturer wants to estimate vaccine demand for a specific disease in the upcoming year.
4. A sudden outbreak of influenza occurs in a specific region, and authorities need to ramp up vaccination efforts.
5. Researchers want to explore the incidence rates of polio in populations with no vaccination coverage.
6. WHO wants to track global progress toward achieving a target of 95% vaccination coverage for measles by 2030.
7. A health agency wants to allocate vaccines to high-risk populations such as children under five and the elderly.
8. A non-profit wants to detect disparities in vaccination coverage across different socioeconomic groups within a country.
9. Authorities want to determine how vaccination rates vary throughout the year.
10. Two regions use different vaccination strategies (e.g., door-to-door vs. centralized vaccination clinics). Authorities want to know which strategy is more effective.

Results:

By the end of this project, learners will achieve:

- A structured SQL database with clean and normalized vaccination and disease data.
- A set of Power BI reports and dashboards that visually represent key insights, trends, and comparisons.
- Insights derived from data analysis, such as vaccination coverage trends, disease outbreaks, and regional disparities.

Project Evaluation metrics:

- **Data Cleaning Process:**
  - Evaluate the handling of missing data, normalization, and consistency checks.
  - Check if the data is ready for analysis in Power BI.
- **SQL Database Quality:**

    - Assess the integrity, normalization, and structure of the SQL database.
    - Ensure proper data types and relationships are defined between tables.
- **Quality of Power BI Visualizations:**
    - Review the clarity and relevance of the Power BI visualizations.
    - Ensure the dashboards are interactive and user-friendly.
- **Insights and Actionability:**
    - Evaluate how well the Power BI reports provide actionable insights for public health policy, resource allocation, and disease prevention strategies.

Technical Tags:

SQL, Power BI, Data Cleaning, Data Analysis, Data Visualization, Healthcare Analytics, Public Health

Data Set:

Source: Vaccination project

Data Set Explanation:

Table 1 : coverage data

Variables:

- **Group:** Categorization of the data. Here, it represents countries.
- **Code:** Unique identifier for the country (ISO Alpha-3 code).

- **Name:** Name of the country.

- **Year :** The year the data is recorded for.
- **Antigen:** Vaccine identifier or code.
- **Antigen_description:** Full description of the vaccine.
- **Coverage_category:** Type of coverage reported, such as administrative or official.
- **Coverage_category description:** Expanded details of the coverage category.
- **Target number:** Number of individuals targeted for vaccination.
- **Dodge:** Number of doses administered.
- **Coverage:** Percentage of the target population that was vaccinated.

Table 2 : Incidence Rate

Variables:

- **Group :** Classification of the data; here, it represents countries.
- **Code:** Unique identifier for the country (ISO Alpha-3 code).
- **Name:** Name of the country.
- **Year:** The year the data is recorded for.
- **Disease:** Code or short name of the disease.
- **Disease description:** Full description of the disease.
- **Denominator:** The population basis used to calculate the incidence rate (e.g., per live births, per total population).
- **Incidence rate:** Number of disease cases per specified population unit.

Table 3 : Reported cases

Variables:

- **Group :** Classification of the data; here, it represents countries.
- **Code:** Unique identifier for the country (ISO Alpha-3 code).
- **Name:** Name of the country.
- **Year:** The year the data is recorded for.
- **Disease:** Code or short name of the disease.
- **Disease description:** Full description of the disease.
- **Cases:** Number of reported cases of the disease for the specified year and region.

Table 4 : Vaccine Introduction

Variables:

- **ISO_3_Code :** Unique 3-letter ISO country code.
- **Country Name:** Name of the country.
- **Who Region:** World Health Organization (WHO) region to which the country belongs.
- **Year:** The year the data is recorded for.
- **Description:** Name of the vaccine or vaccine type.

- **Intro:** Indicates whether the vaccine has been introduced into the country's vaccination program.

Table 5 : Vaccine Schedule Data

Variables:

- **ISO_3_Code :** Unique 3-letter ISO country code.
- **Country Name:** Name of the country.
- **Who Region:** World Health Organization (WHO) region to which the country belongs.
- **Year:** The year the data is recorded for.
- **Vaccine code:** Code for the vaccine.
- **Vaccine description:** Name and details of the vaccine.
- **Schedule rounds:** The dose or round of the vaccine in the schedule.
- **Target pop:** Specific population targeted for the vaccine dose.
- **Target pop description:** Detailed description of the target population.
- **Georarea:** Geographic area of administration.
- **Age administered:** Age or time of vaccine administration.
- **Source comment:** Additional information or context for administration.

Project Deliverables:

Source Code:

- Python scripts for data extraction and cleaning.
- SQL queries for creating and populating tables in the database.

SQL Database:

- Structured and normalized database containing the cleaned data.

Power BI Reports:

- A set of interactive reports and dashboards showcasing key insights.

Documentation:

- Detailed documentation explaining the process, challenges faced, and solutions implemented.

Project Guidelines:

- Use best practices for SQL database design and normalization.
- Follow Power BI best practices for creating interactive, user-friendly dashboards.

Timeline: 7 Days

Check your mail for the submission deadline of the project.